

Improving Accuracy of Goodness-of-fit Test.

Kris Duszak and Jan Vrbik
Brock University

October 28, 2014

Abstract

It is well known that the approximate distribution of the usual test statistic of a goodness-of-fit test is chi-square, with degrees of freedom equal to the number of categories minus 1 (assuming that no parameters are to be estimated – something we do throughout this article). Here we show how to improve this approximation by including two correction terms, each of them inversely proportional to the total number of observations.

1 Goodness-of-fit Test: A Brief Review

To test whether a random independent sample of size n comes from a specific distribution can be done by dividing all possible outcomes of the corresponding random variable (say U) into k distinct regions (called CATEGORIES) so that these have similar probabilities of happening. The sample of n values of U is then converted into the corresponding observed frequencies, one for each category (we denote these X_1, X_2, \dots, X_k), equivalent to sampling a multinomial distribution with probabilities p_1, p_2, \dots, p_k (computed, for each category, based on the original distribution). The new random variables X_i have expected values given by $n \cdot p_i$ (where i goes from 1 to k) and variance-covariance matrix given by

$$n \cdot (\mathbb{P} - \mathbf{p} \mathbf{p}^T)$$

where \mathbf{p} is a column vector with k elements (the individual p_i probabilities), and \mathbb{P} is similarly an $k \times k$ *diagonal* matrix, with the same p_i probabilities on its main diagonal.

The usual test statistic is

$$T = \sum_{i=1}^k \frac{(X_i - n \cdot p_i)^2}{n \cdot p_i} \equiv \sum_{i=1}^k Y_i^2 \quad (1)$$

where

$$Y_i \equiv \frac{X_i - n \cdot p_i}{\sqrt{n \cdot p_i}} \quad (2)$$

equivalent to (in its vector form)

$$\mathbf{Y} = \frac{\mathbb{P}_k^{-1/2}(\mathbf{X} - n \cdot \mathbf{p})}{\sqrt{n}} \quad (3)$$

where \mathbf{X} is a column vector of the X_1, X_2, \dots, X_k observations.

The Y_i 's have a mean of zero and their variance-covariance matrix is

$$\mathbb{V} = \mathbb{P}^{-1/2}(\mathbb{P} - \mathbf{p} \mathbf{p}^T) \mathbb{P}^{-1/2} = \mathbb{I} - \mathbf{p}^{1/2}(\mathbf{p}^{1/2})^T \quad (4)$$

where \mathbb{I} is the $k \times k$ unit matrix and $\mathbf{p}^{1/2}$ denotes a column vector with elements equal to $p_1^{1/2}, p_2^{1/2}, \dots, p_k^{1/2}$. The matrix (4) is *idempotent*, since

$$\mathbf{p}^{1/2}(\mathbf{p}^{1/2})^T \mathbf{p}^{1/2}(\mathbf{p}^{1/2})^T = \mathbf{p}^{1/2}(\mathbf{p}^{1/2})^T$$

and its *trace* is $k - 1$, since

$$\text{Tr} \left[\mathbf{p}^{1/2}(\mathbf{p}^{1/2})^T \right] = \text{Tr} \left[(\mathbf{p}^{1/2})^T \mathbf{p}^{1/2} \right] = \sum_{i=1}^k p_i = 1.$$

Because the k -dimensional distribution of (3) tends (as $n \rightarrow \infty$) to a Normal distribution with zero means and variance-covariance matrix of (4), (1) must similarly converge to the χ_{k-1}^2 distribution (assuming that U does have the hypothesized distribution). A substantial disagreement between the observed frequencies X_i and their expected values $n \cdot p_i$ will be reflected by the test statistic T exceeding the (right-hand-tail) critical value of χ_{k-1}^2 , leading to a rejection of the null hypothesis.

Since the sample size is always finite, the critical value (computed under the assumption that $n \rightarrow \infty$) will have an error roughly proportional to $\frac{1}{n}$. To remove this error is an objective of this article.

2 $\frac{1}{n}$ proportional correction

A small modification of the results of [1] indicate that a substantially better approximation (which removes the $\frac{1}{n}$ -proportional error) to the probability density function (PDF) of the distribution of T (under the null hypothesis) is

$$\begin{aligned} & \chi_{k-1}^2(t) \cdot \left(1 + B \cdot \left(\frac{t^2}{(k-1)(k+1)} - \frac{2t}{k-1} + 1 \right) + \right. \\ & \left. C \cdot \left(\frac{t^3}{(k-1)(k+1)(k+3)} - \frac{3t^2}{(k-1)(k+1)} + \frac{3t}{k-1} - 1 \right) \right) \end{aligned} \quad (5)$$

where $\chi_{k-1}^2(t)$ is the PDF of the regular chi-square distribution and

$$B = \frac{1}{8} \sum_{i,j=1}^k \kappa_{i,i,j,j} \quad (6)$$

$$C = \frac{1}{8} \sum_{i,j,\ell=1}^k \kappa_{i,j,j} \kappa_{i,\ell,\ell} + \frac{1}{12} \sum_{i,j,\ell=1}^k \kappa_{i,j,\ell}^2 \quad (7)$$

where $\kappa_{i,j,\ell}$ and $\kappa_{i,j,\ell,h,,}$ are cumulants of the (multivariate) \mathbf{Y} distribution. They can be found easily, based on the logarithm of the joint moment generating function of (2), namely

$$M = n \cdot \ln \left(\sum_{m=1}^k p_m \exp \left(\frac{t_m}{\sqrt{n \cdot p_m}} \right) \right) - \sum_{m=1}^k t_m \sqrt{n \cdot p_m}$$

by differentiating M with respect to t_i , t_j and t_ℓ to get $\kappa_{i,j,\ell}$ (and the extra t_h to get $\kappa_{i,j,\ell}$), followed by setting all $t_m = 0$.

This yields

$$\begin{aligned} \kappa_{i,i,i} &= \frac{(1-p_i)(1-2p_i)}{\sqrt{n \cdot p_i}} \\ \kappa_{i,i,j} &= -\frac{\sqrt{p_j}(1-2p_i)}{\sqrt{n}} \\ \kappa_{i,j,\ell} &= \frac{2\sqrt{p_i \cdot p_j \cdot p_\ell}}{\sqrt{n}} \end{aligned}$$

and

$$\begin{aligned} \kappa_{i,i,i,i} &= \frac{(1-p_i)(1-6p_i+6p_i^2)}{n \cdot p_i} = \frac{1}{n} \left(\frac{1}{p_i} - 7 + 12p_i - 6p_i^2 \right) \\ \kappa_{i,i,j,j} &= \frac{2p_i + 2p_j - 6p_i \cdot p_j - 1}{n}. \end{aligned}$$

Using these formulas, we can proceed to compute

$$\begin{aligned} B &= \frac{1}{8} \sum_{i=1}^k \kappa_{i,i,i,i} + \frac{1}{8} \sum_{i \neq j}^k \kappa_{i,i,j,j} = \\ &= \frac{1}{8n} (Q - 7k + 12s_1 - 6(s_1^2 - 2s_2) + 2(k-1)s_1 + 2(k-1)s_1 - 12s_2 - k(k-1)) \end{aligned}$$

where

$$Q \equiv \sum_{i=1}^k \frac{1}{p_i}$$

and s_1 and s_2 are the first two elementary symmetric polynomials in p_i , i.e.

$$\begin{aligned} s_1 &= \sum_{i=1}^k p_i \\ s_2 &= \sum_{i < j}^k p_i \cdot p_j \end{aligned}$$

(note that $\sum_{i=1}^k p_i^2 = s_1^2 - 2s_2$). Realizing that $s_1 = 1$, the expression for B can be simplified to

$$B = \frac{1}{8n} (Q - k^2 - 2k + 2). \quad (8)$$

When choosing the categories in a manner which makes all p_i equal to $1/k$, the last expression reduces to

$$-\frac{k-1}{4n}$$

Similarly,

$$\begin{aligned} C &= \frac{1}{8} \sum_{i=1}^k \kappa_{i,i,i}^2 + \frac{1}{4} \sum_{i \neq j}^k \kappa_{i,i,i} \kappa_{i,j,j} + \frac{1}{8} \sum_{i \neq j}^k \kappa_{i,j,j}^2 + \frac{1}{8} \sum_{i \neq j \neq \ell}^k \kappa_{i,j,j} \kappa_{i,\ell,\ell} \\ &+ \frac{1}{12} \sum_{i=1}^k \kappa_{i,i,i}^2 + \frac{1}{4} \sum_{i \neq j}^k \kappa_{i,i,i}^2 + \frac{1}{12} \sum_{i \neq j \neq \ell}^k \kappa_{i,j,\ell}^2 \\ &= \frac{5}{24n} \sum_{i=1}^k \frac{(1-p_i)^2(1-2p_i)^2}{p_i} - \frac{1}{4n} \sum_{i \neq j}^k (1-p_i)(1-2p_i)(1-2p_j) \\ &+ \frac{3}{8n} \sum_{i \neq j}^k p_j(1-2p_i)^2 + \frac{1}{8n} \sum_{i \neq j \neq \ell}^k p_i(1-2p_j)(1-2p_\ell) + \frac{1}{3n} \sum_{i \neq j \neq \ell}^k p_i p_j p_\ell \\ &= \frac{5}{24n} \sum_{i=1}^k \left(\frac{1}{p_i} - 6 + 13p_i - 12p_i^2 + 4p_i^3 \right) \\ &- \frac{1}{4n} \sum_{i=1}^k (k(1-3p_i+2p_i^2) - 3 + 11p_i - 12p_i^2 + 4p_i^3) \\ &+ \frac{9}{24n} \sum_{i=1}^k (1-5p_i+8p_i^2-4p_i^3) + \frac{1}{8n} \sum_{i \neq j \neq \ell}^k p_i(1-2p_j-2p_\ell) + \frac{5}{6n} \sum_{i \neq j \neq \ell}^k p_i p_j p_\ell \\ &= \frac{1}{24n} (5Q - 21k + 20 + 12(s_1^2 - 2s_2) - 16(s_1^3 - 3s_1s_2 + 3s_3)) \\ &- \frac{1}{4n} (k(k-3+2(s_1^2 - 2s_2)) - 3k + 11 - 12(s_1^2 - 2s_2) + 4(s_1^3 - 3s_1s_2 + 3s_3)) \\ &+ \frac{1}{8n} ((k-2)(k-1) - 2(k-2)2s_2 - 2(k-2)2s_2) + \frac{5}{n} s_3 \\ &= \frac{1}{24n} (5(Q - k^2) + 2(k-1)(k-2)) \end{aligned}$$

where

$$s_3 = \sum_{i < j < \ell}^k p_i \cdot p_j \cdot p_\ell$$

Note that

$$\sum_{i=1}^k p_i^3 = s_1^3 - 3s_1s_2 + 3s_3$$

and that the final formula reduces to

$$C = \frac{(k-1)(k-2)}{12n}$$

in the case of all categories being equally likely.

The corresponding *distribution function* is given by

$$F_T(u) = \int_0^u \chi_{k-1}^2(t) dt - 2\chi_{k-1}^2(u) \cdot \frac{u}{k-1} \cdot \left[B \cdot \left(\frac{u}{k+1} - 1 \right) + C \cdot \left(\frac{u^2}{(k+1)(k+3)} - \frac{2u}{k+1} + 1 \right) \right] \quad (9)$$

which can be used for a substantially more accurate computation of critical values of T (by setting $F_T(u) = 1 - \alpha$ and solving for u).

3 Monte Carlo Simulation

We investigate the improvement achieved by this correction by selecting (rather arbitrarily) the value of k (from the most common 5 to 15 range), the individual components of \mathbf{p} , and the sample size n (with a particular interest in small values). Then we generate a million of such samples and, for each of these, compute the value of T . The resulting empirical (yet ‘nearly exact’) distribution is summarized by a histogram, which is then compared with the χ_{k-1}^2 approximation, first *without* and then *with* the proposed correction of (5). Marginally we mention that, when $p_i = \frac{1}{k}$ for all i (the *uniform* case), the set of potential values of T becomes rather small (the values range from $k - n$ to $n(k - 1)$ in steps of $2k/n$). For large enough n , the shape of the exact distribution still follows the χ_{k-1}^2 curve, but in a correspondingly ‘discrete’ manner. Our examples tend to avoid this complication by making the p_i values sufficiently distinct from each other; the exact T distribution remains discrete, but the number of its possible values increases so dramatically that this is no longer an issue (unless n is extremely small, the distribution can be considered, for any practical purposes, to be continuous).

The simulation reveals that, when $k = 5$, the essential discreteness of the T distribution remains ‘visible’ (even with a *non-uniform* choice of p_i s) unless n is at least 20. Such a relatively large value of n (an average of 4 per category) results in only a marginal improvement achieved by our correction – see Fig. 1, with the blue curve being the basic χ_{k-1}^2 approximation and the red one representing (5).

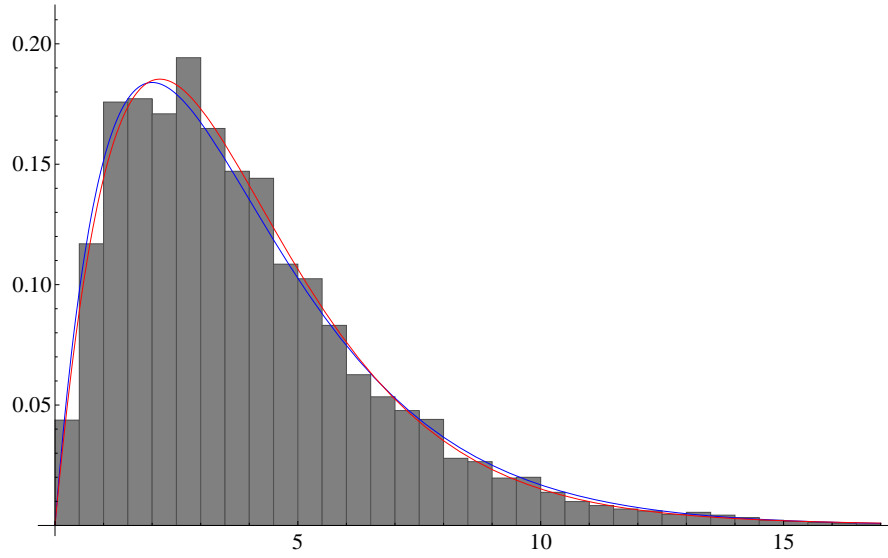


FIGURE 1.

When $k = 10$ and the \mathbf{p} values are reasonable ‘diverse’ (those of our example range from 0.033 to 0.166), the discreteness of the exact T distribution is less of a problem (even though still showing – see Fig. 2), even for n as low as 12 (our choice). The new formula already proves to be a definite improvement over the basic approximation:

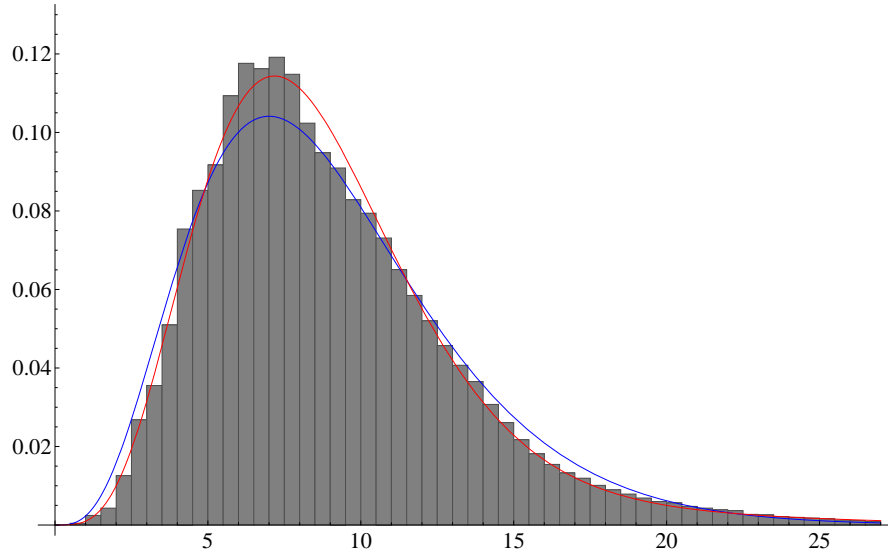


FIGURE 2.

Finally, when $k = 15$, the distribution becomes almost perfectly smooth (eliminating all traces of discreteness – see Fig. 3) even for $n = 10$. Unfortu-

nately, this sample size is now so small that it is our approximation itself which starts showing a visible error (for this value of k , this happens whenever the absolute value of either B or C exceeds 2.25; in this example $B = 0.31$ and $C = 2.62$). The general rule of thumb is that neither B nor C should exceed $0.15k$ (beyond that, the approximation may become increasingly nonsensical).

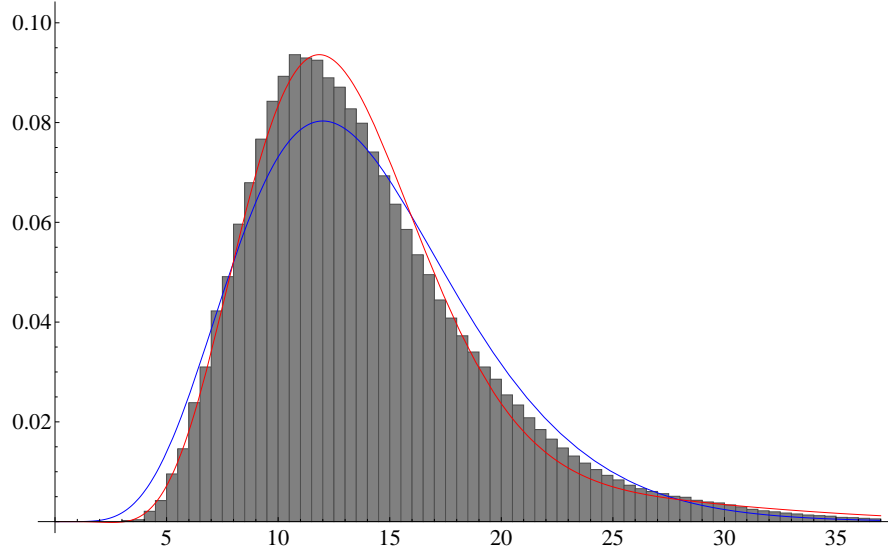


FIGURE 3.

To demonstrate the true superiority of the new approximation, we now use $k = 15$ and $n = 15$, with the individual probabilities ranging from 0.028 to 0.116 (Fig. 4). Since now $B = 0.085$ and $C = 1.54$, the new approximation (unlike the old one, which is clearly off the mark) represents a decent agreement with the ‘exact’ answer.

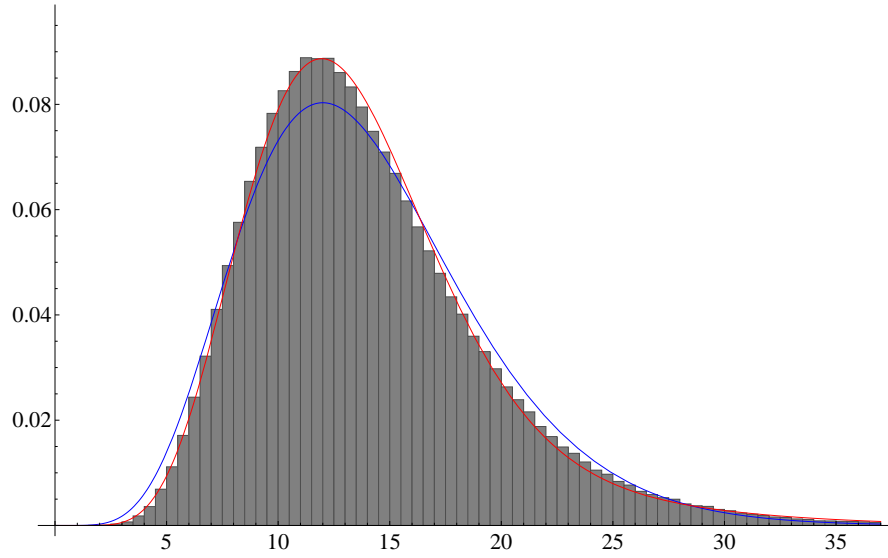


FIGURE 4.

4 Conclusion

Using the χ^2 approximation to perform the usual goodness-of-fit test, the number of observations should be as large as possible; when this becomes impractical (e.g. each observation is very costly), one can still achieve good accuracy by:

1. increasing the number of categories (one should aim for the 10 – 15 range); this inevitably results in reducing the average number of observations per category – in spite of that, the test becomes more accurate,
2. choosing categories in such a way that their individual probabilities are all distinct from each other (avoiding the $p_i = 1/k$ situation) but, at the same time, not letting any one of them become too small (this would increase, often dramatically, the value of each B and C of our correction – see the next item),
3. using the $\frac{1}{n}$ proportional correction of (9), but monitoring the values of B and C (neither of them should be bigger, in absolute value, than $0.15k$).

References

- [1] Vrbik J: “Accurate Confidence Regions based on MLEs” *Advances and Applications in Statistics* **32** #1 (2013) 33-56